Text Independent Speaker Identification Using Gaussian Mixture Model

Chee-Ming Ting, Sh-Hussain Salleh, Tian-Swee Tan, A. K. Ariff.
Center for Biomedical Engineering, Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81300 Skudai, Johor, Malaysia.

***Abstract*— This paper describes text-independent (TI) Speaker Identification (ID) using Gaussian mixture models (GMM). The use of GMM approach is motivated by that the individual Gaussian components of a GMM are shown to represent some general speaker-dependent spectral shapes that are effective for speaker identity modeling. For speaker model training, a fast re-estimation algorithm based on highest likelihood mixture clustering is introduced. In this work, the GMM is evaluated on TI Speaker ID task via series of experiments (model convergence, effect of feature set, number of Gaussian components, and training utterance length on identification rate). The database consisted of Malay clean sentence speech database uttered by 10 speakers (3 female and 7 male). Each speaker provides the same 40 sentences utterances (average length- 3.5s) with different text. The sentences for testing were different from those for training. The GMM achieved 98.4% identification rate using 5 training sentences. The model training based on highest likelihood clustering is shown to perform comparably to conventional expectation-maximization training but consumes much shorter computational time.**

***Keywords*— Speaker Indentification, Gaussian Mixture Model**

I. Introduction

There are numerous measurements and signals have been investigated for the use in biometric systems. Each has its advantages and disadvantages in term of accuracy and deployment. Speech signal, which conveys information about the identity of the speaker, has become compelling in biometric area. The advantage of voice biometrics is that firstly, speech is a natural signal to produce. Secondly, speech has been used as common communication medium in many applications (e.g. telephone transaction), so users can provide the signals conveniently. Besides, speech data acquisition is low-cost, using microphone and existing telephone systems without much extra installations.

Depending on applications, speaker recognition is generally divided into two tasks: verification (verifying if a person is whom he/she claims by determining whether the input voice is from that particular person) and identification (select the correct speaker out of a given population, who is best matched to the input voice sample). Speaker verification (SV) is generally more important than speaker identification (SI) for most commercial applications. Besides, the speech input can be constrained to be a known phrase or 'password' (text-dependent (TD) systems). These systems consider the temporal information of fixed text and are capable of improving the accuracy. However TD methods may be inconvenient since the password has to be remembered. This method also cannot be used when the speaker is uncooperative and when the verification is required during the normal conversation. To avoid these problems, a more flexible system which able to operate without explicit user cooperation and independent of spoken utterances (text-independent (TI)) is needed. This paper focuses on text-independent speaker identification.

Many approaches have been proposed for TI speaker recognition. First is the VQ based method which uses VQ codebooks as an efficient means of characterizing speaker-specific feature [1]. An input utterance is first vector-quantized using the codebook of each reference speaker, and the VQ distortion is used for making recognition decision. To better modeling the acoustic feature and incorporate the temporal structure modeling, the Hidden Markov Models (HMM) have been used as probabilistic speaker model for both TI and TD tasks. Poritz [2] proposed a five state ergodic HMM, which classify acoustic events into broad phonetic categories corresponding to HMM states, to characterize each speaker in TI task. However, Matsui [3] found that TI performance was unaffected by discarding transition probabilities in HMM models.

Rose and Reynolds [4] introduced a methods based on Gaussian Mixture Model (GMM) (corresponds to a single state continuous ergodic HMM by [3]) to model speaker identity. The GMM, on the other hand, provide probabilistic model of the underlying acoustic properties of a person but do not impose any Markovian constraints between the acoustic classes by discarding the transition probabilities in the HMM models. The use of GMM for speaker identity modeling is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixture to model arbitrary densities. The GMM has been firstly used for TI speaker identification [5] and is extended to speaker verification on several publicly available speech corpora [6]. The GMM was also shown to outperform the conventional Vector Quantization (VQ) method and discriminative method (Radial Basis Function) in TI speaker ID task [5].

In this work, GMM approach is investigated for TI speaker identification. In speaker model training, expectation-maximization EM re-estimation algorithm can be used for estimating the GMM parameters [5,6]. This paper applies a re-estimation algorithm based on highest mixture likelihood clustering on speaker identification. This algorithm has been

1-4244-1355-9/07/$25.00 @2007 IEEE

proposed by the author [7] to train the GMM models and its evaluation on TI- speaker verification task gives convincing result on verification rate and model convergence. This paper extends the work to speaker identification. The algorithm is an iterative process of two steps: (1) cluster the training vectors to the mixture component with the highest likelihood and (2) re-estimate parameters of each component. This process increases the likelihood of training data at each iteration. This algorithm performs comparably to EM training with much shorter computational time, which will shorten the enrollment time.

The paper is organized as follows. Section 2 describes the GMM based speaker identification system. Section 3 present the experimental evaluation on the GMM model in term of model convergence, effect of feature set, number of Gaussian components, and training utterance length on identification rate. Conclusion is given in the last section.

II. GMM Speaker Identification System

The GMM speaker identification system consists of the following elements: speech processing, Gaussian mixture model, parameter estimation of GMM speaker model, and identification.

A. Speech Processing

Most of the front end used in SV systems relies on cepstral representation of speech. The Mel-scale Frequency cepstral coefficients (MFCC) extraction is used in front-end processing. The sampled speech signals are pre-emphasized with filter. Then, the waveform is blocked into 15ms-width frames with 5ms frame rate. Each frame is multiplied by a Hamming window. In the process of MFCC extraction, the DFT spectrums are filtered by triangular windows, which are arranged in Mel-scale (designed to approximate the amplitude frequency response of human ear). Next, log compression is put on the output of each filter. Finally, Discrete Cosine Transform (DCT) is applied to de-correlate feature vector of MFCCs.

B. Gaussian Mixture Model

The Gaussian mixture density used for the likelihood function is a weighted linear combination of *M* uni-model Gaussian component densities, defined as follows:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i b_i(\vec{x}) . \qquad (1)$$

,where $\vec{x}$ is a *D*-dimensional vector, $b_i(\vec{x}), i = 1, \ldots, M$ are the component densities and $w_i, i = 1, \ldots, M$ are the mixture weights. Each component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\Pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \qquad (2)$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weight satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$.

Collectively, the parameters of the Gaussian mixture density model are denoted as $\lambda = (w_i, \vec{\mu}_i, \Sigma_i), i = (1, \ldots, M)$. Each speaker is represented by a GMM and is referred to by his/her model (speaker specific model). There are two form of Σ, namely diagonal matrices and full covariance matrices .Use of diagonal matrices consume less training data and time while outperforming empirically full covariance matrices.

C. GMM Parameters Estimation

Given training speech, speaker model training is to estimate the GMM parameters via maximum likelihood (ML) estimation, which maximizes the likelihood of the GMM. For a T training vectors pattern *X*, the likelihood is

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda) \qquad (3)$$

The ML estimate can be obtained using iterative expectation-maximization (EM) algorithm. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for a given X, i.e. for iterations *k* and *k*+1, $p(X \mid \lambda^{(k+1)}) > p(X \mid \lambda^{(k)})$. The EM equations for GMM training given as follows can be found in [5].

$$\overline{w}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda) \qquad (4)$$

$$\bar{\vec{\mu}}_i = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_i, \lambda) \vec{x}_t}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} \qquad (5)$$

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i \mid \vec{x}_i, \lambda) x_t^2}{\sum_{t=1}^{T} p(i \mid \vec{x}_t, \lambda)} - \overline{\mu}_i^2 \qquad (6)$$

,where σ_i^2, x_t, and μ_i refer to arbitrary elements of the vectors σ_i^2, $\vec{x}_t$, and $\vec{\mu}_i$ respectively. $p(i \mid \vec{x}_t, \lambda)$ is a posteriori probability for acoustic class *i.* The iteration is repeated until some convergence threshold is reached.

Careful selection of model initialization is important for the optimal model convergence. In this work, VQ clustering (involves using Euclidean distortion measures and VQ design algorithm –Linde-Bunzo-Gray (LBG) algorithm) is for initialization. The LBG algorithm clusters the training vectors into a set of *M* clusters. The vectors in each of the VQ-cluster *i* are used to estimate the corresponding m Gaussian mixture components.

This paper proposes an alternative GMM re-estimation algorithm based on highest mixture component likelihood clustering. The iterative algorithm meets the maximum likelihood estimation criterion as the EM algorithm. The algorithm consists of two steps as shown in Figure 1: (1) cluster the training vectors to the mixture component with the

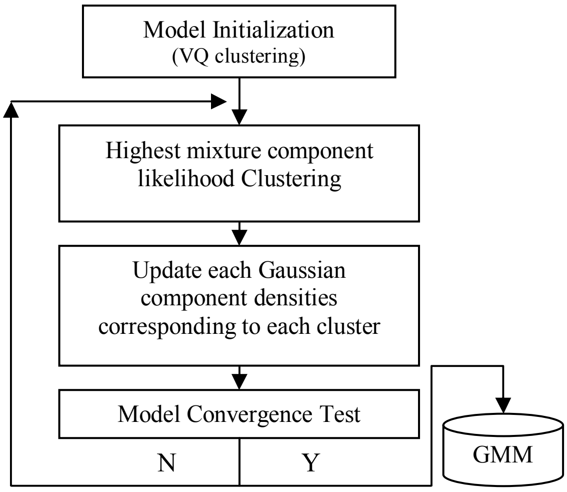
Model Initialization
(VQ clustering)
Highest mixture component likelihood Clustering
Update each Gaussian component densities corresponding to each cluster
Model Convergence Test
N
Y
GMM

Figure 1. GMM re-estimation based on highest mixture component likelihood clustering.

highest likelihood, and (2) re-estimate parameters of each component. Based on the initialized model, each training vector is clustered into regions $C_i, i = (1,\ldots,M)$, *M* is the number of mixture components (model order) of GMM. The clustering criterion is associating each training vector to the Gaussian mixture components *i* with highest likelihood, such that

$$C_i = \arg\max_{1\le i\le M} b_i(\vec{x}) \tag{7}$$

Each cluster represents one of the M mixture densities. Next, the vectors in each of the cluster *i* are used to estimate the corresponding i^{th} Gaussian mixture components using simple averaging estimation derived as follows:

$\overline{w}_i$ = number of vectors classified in cluster *i* / total number of training vectors.

$\overline{\vec{\mu}}_i$ = sample mean of vectors classified in cluster *i*.

$\overline{\Sigma}_i$ = sample covariance matrix of vectors classified in cluster *i*.

Iteration is repeated until certain convergence criterion is met.

D. Speaker Identification

For a close set speaker identification, a group of speakers *S* with each speaker represented by speaker specific GMM's $\lambda_1, \lambda_2, \ldots, \lambda_T$. Let a speech signal uttered by an unknown speaker, which after front end processing, gives a feature vector pattern, *X*. The feature is classified to the speaker $\hat{S}$, whose model likelihood is the highest, in formal term [5],

$$\hat{S} = \arg\max_{1\le k\le S} p(X \mid \lambda_k). \tag{8}$$

Assume independence between observation vectors the above can be formulated in logarithmic term,

$$\hat{S} = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \tag{9}$$

$p(\vec{x}_t \mid \lambda_k)$ is given by (1).

III. Experimental Evaluation

A. Database and Experiment Conditions

The database for system evaluation consists of phonetically balanced Malay sentences utterances by 7 male and 3 female client speakers with each provides the same 40 sentences utterances with different text. This database was recorded on one session in the same recording room with same microphone for all speakers for all sessions. The average sentences duration is approximately 3.5 s. A subset of sentences is used for training the speaker specific model. The other subset is used for testing. The training sentences with different text are same for all speakers. The testing sentences were different from those for training but same for all speakers. For identification, an unknown speech signal which has been transformed into MFC feature pattern is classified into speaker whose GMM model gives highest likelihood. A series of experiments were established to evaluate the systems.

B. Performance Comparison between EM & Highest Mixture Likelihood Clustering Training.

Performance comparison between speaker model training using EM and highest mixture likelihood clustering is investigated in term of model convergence and identification rate.

A speaker model with number of Gaussian components 16, and 16 dimensional MFCCs was trained using both training algorithm to investigate their convergence properties and consumed computational time. 20 utterances were used for training. The Figure 2 shows the convergence rate of the two training algorithms through the total log likelihood per frame of the training data at each training iteration. The iteration stops when $|\,p(X \mid \lambda^{(k+1)}) - p(X \mid \lambda^{(k)})\,| < 0.03$ .The EM training gives more optimal convergence at each iteration than the highest likelihood clustering and achieves a higher local

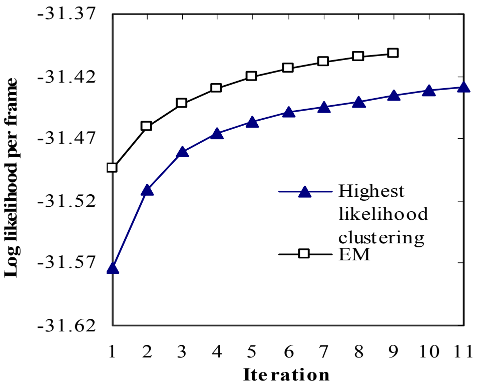

Figure 2. : Convergence rate of the EM and highest likelihood clustering re-estimation algorithm.

TABLE I. TRAINING DURATION OF EM AND HIGHEST LIKELIHOOD CLUSTERING TRAINING.

Training algorithm	Training Duration
EM	9 min
Highest likelihood clustering	1 min 6 s

TABLE II. COMPARISON OF EM AND HIGHEST LIKELIHOOD CLUSTERING TRAINING ON IDENTIFICATION RATE.

Training algorithm	Identification Rate
EM	95.18%
Highest likelihood clustering	94.80%

maxima. The duration of each iteration for EM and highest likelihood clustering are 6s and approximate 1min respectively. Table 1 shows the training duration of both algorithms. The highest likelihood clustering consumes less training time than EM training. These results have been reported in [7].

Table 2 shows the comparison between EM and highest likelihood clustering training on identification rate. 10 sentences were used for training and 25 sentences were used for testing. Two set of speaker models with 4 Gaussian components were trained using 8 iterations of both training algorithms respectively. The performance of the highest likelihood clustering is comparable with the EM training but consume less computational time.

C. Effect of Different Number of Gaussian Mixture Components and Amount of Training Data

No objective way to determine the correct number of mixture components (model order) and the model dimension *a priori*. For saving the identification time, the objective is to choose the minimum number of components necessarily for adequate speaker modeling. However, too few components will not be able to accurately model the distinguished characteristics of a speaker distribution. Too many components relative to limited training data induce too many free parameters to be estimated reliably, thus degrade performance. Besides, small amount of training data is crucial to facilitate client enrolment to the system, with the trade-off that the insufficient data unable to train the model reliably.

The following experiment investigates the effect of different number of Gaussian mixture components on identification rate for different amount of training data. MFCC feature dimension is fixed to 12. The speaker models with model order varied from 1 to 32, were trained using 5, 10, and 15 training sentences. 25 sentences of different text from the training set were used for testing. Figure 2 shows the identification results.

Generally, for all model order, increasing the amount of training data increases the identification rate. For all amount of training data, there is a sharp increase in performance from 1 to 4 components, and start leveling off at 8 components. Identification rate peaks at 16 components. This shows that at least 8 components are sufficient to fit different acoustic categories, and gives better discriminating power to yield high performance, for one speaker model. Compared to the relatively constant performance, the performance for the small amount of training data (5 sentences) drops at 32 mixture components. This is because there are too many parameters to be estimated reliably with relatively insufficient training data.

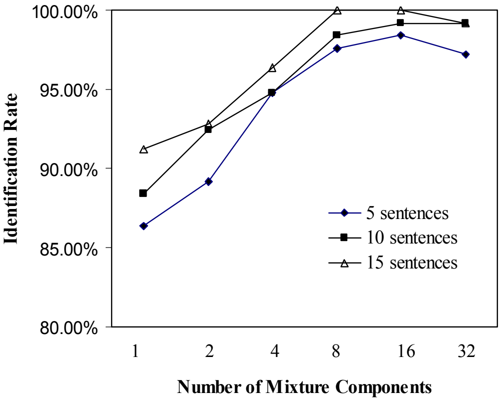

Figure 3. Speaker identification rate versus number of mixture components for different amount of training data.

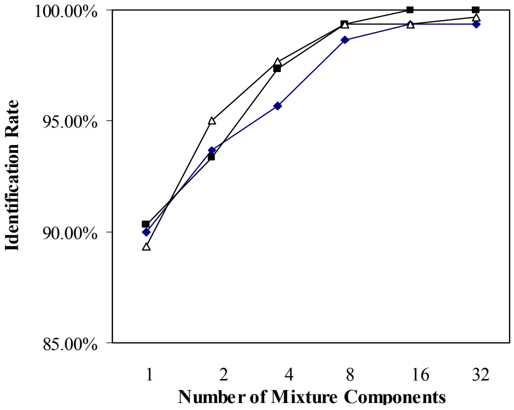

Figure 4. Speaker identification rate versus number of mixture components for different feature sets.

D. Effect of Feature Set on Performance forDifferent Number of Gaussian Mixture Components

The cepstrum difference coefficients are widely used in speaker recognition [5, 8]. The use of difference coefficients is motivated by its ability to capture dynamic or transitional cepstral information. Besides, for speaker recognition this feature sets contain channel invariant feature and speaker specific information. The difference coefficients are called dynamic feature while cepstral coefficients are called static feature. The difference coefficients were tested for channel compensation for telephone speech on TI speaker identification task, and shows improvements [5]. The difference coefficients have been shown when used by themselves; do not perform as well as static feature. They are normally used in combination with static feature by being appended to the static feature vectors.

The following experiments investigate the use of combined static and dynamic MFCC features. Combination with first and second order difference coefficients was tested. Δ and ΔΔ denotes the first and second order difference coefficient respectively. 12 dimensional static features were used. 10

sentences were used for training and 30 sentences were used for testing. Figure 4 shows the speaker identification rate versus number of mixture components for different feature sets. As expected increasing mixture components increase the identification rate. The incorporation of first order differences coefficients to static outperforms the use of static feature by themselves for all number of mixture components. However, the incorporation of the first and second order coefficient gain not much advantage and even slightly decrease the performance with the increase of mixture components, compared to merely first order coefficient incorporation. This is may be due to the increasing feature parameters dimension made the reliable model estimation difficult. These results postulate that the use of first order difference coefficients is sufficient to capture the transitional information while maintaining reasonable dimensional complexity.

IV. Conclusion

A GMM based text-independent speaker identification system has been described. This paper extends the use of highest mixture likelihood clustering to speaker identification. The alternative GMM training algorithm performs comparably to conventional EM training but with less computational time. Increasing the amount of training data increases the identification rate. Experimental result shows that increasing the mixture components of the speaker model improves the performance, limited by amount of training data. The use of first order difference coefficients is sufficient to capture the transitional information with reasonable dimensional complexity. The 12 dimensional 16 order GMM trained with highest mixture likelihood clustering using 5 training sentences achieved 98.4% identification rate.

Acknowledgment

This research project is supported by CBE (Center of Biomedical Engineering) at Universiti Teknologi Malaysia and funded by Minister of Higher Education (MOHE), Malaysia under grant “Support Vector Machines (SVM) As a Large Margin Classifier For Machine Learning Algorithm “ Vot 78029.

References

[1] Rosenberg A. E. & Soong F. K.. “Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes,” In Proc. ICASSP, pp 873-876, 1986.

[2] Poritz A. B., ”Linear predictive Hidden Markov Model and speech signal,” In Proc. ICASSP, pp 1291-1294, 1982.

[3] Matsui T. & Furui S., “Comparison of Test-Independent Speaker Recognition Methods Using VQ Distortion and Discrete/Continuous HMMs,” In Proc. ICASSP, pp 157-160, 1992.

[4] Rose R. & Reynolds D. A., “Text-Independent Speaker Identification Using Automatic Acoustic Segmentation” Proc. ICASSP pp 293-296, 1990.

[5] Reynolds D. A. & Rose R., “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Model,” IEEE Trans. Speech and Audio Processing, vol. 3, pp. 72-83, 1995.

[6] Reynolds D. A. *et.. al.*, “Speaker Verification Using Adapted Gaussian Mixture Models.” Digital Signal Processing, vol. 10 pp.19-41, 2000.

[7] C.-M. Ting, Sh-Hussain Salleh, Ariff A. K., “Text-Independent Speaker Verification using Gaussian Mixture Model Approach,” Asian Biometrics Consortium Conference and Exhibition, 2007. (To be published).

[8] Soong, F., and Rosenberg, A., “ On the Use of Instantaneos and Transitional Spectral Information in Speacker Recognation,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 36, pp. 871-879, June 1988.